

Knowledge Discovery and Data Mining

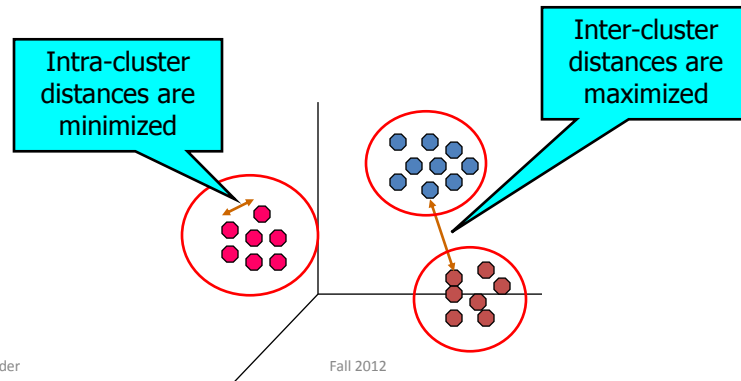
Unit # 11

Acknowledgement

- Most of the slides in this presentation are taken from course slides provided by
 - Han and Kimber (Data Mining Concepts and Techniques) and
 - Tan, Steinbach and Kumar (Introduction to Data Mining)
 - Several other online sources

Cluster Analysis

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



Sajjad Haider

Fall 2012

3

Cluster Analysis (Cont'd)

- Cluster: a collection of data objects
 - Similar to one another within the same cluster
 - Dissimilar to the objects in other clusters
- Cluster analysis
 - Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters
- Unsupervised learning:** no predefined classes
- Typical applications
 - As a **stand-alone tool** to get insight into data distribution
 - As a **preprocessing step** for other algorithms

Sajjad Haider

Fall 2012

4

Google News

iPhone activation headaches still trouble users

Computerworld - 1 hour ago

July 02, 2007 (Computerworld) -- It took Iain Gillott 47 hours to activate his iPhone after waiting in the Texas heat Friday afternoon to buy one.

Most iPhone users thrilled but a few are iRate Reuters

Apple iPhone Arrives in the US Techtree.com

Forbes - ZDNet - Ars Technica - Wired News

[all 562 news articles »](#)



Local10.com

- They didn't pick all 3,400,217 related articles by hand...
- Or Amazon.com
- Or Netflix...

McCain Considers Ways to Reshape Campaign

Washington Post - 35 minutes ago

By Alec MacGillis Sen. John McCain's presidential campaign today announced widespread cutbacks and said it was considering whether to accept public campaign funds after another disappointing fundraising effort that has left the Arizona Republican with ...

McCain's Troubles Mount New York Times

McCain Campaign Struggling, Reduces Staff ABC News

CBS News - Reuters - Angus Reid Global Monitor - Sarasota Herald-Tribune

[all 291 news articles »](#)



Seattle Post
Intelligencer

Sajjad Haider

Fall 2012

5

Examples of Clustering Applications

- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- Land use: Identification of areas of similar land use in an earth observation database
- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost
- City-planning: Identifying groups of houses according to their house type, value, and geographical location
- Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults

Sajjad Haider

Fall 2012

6

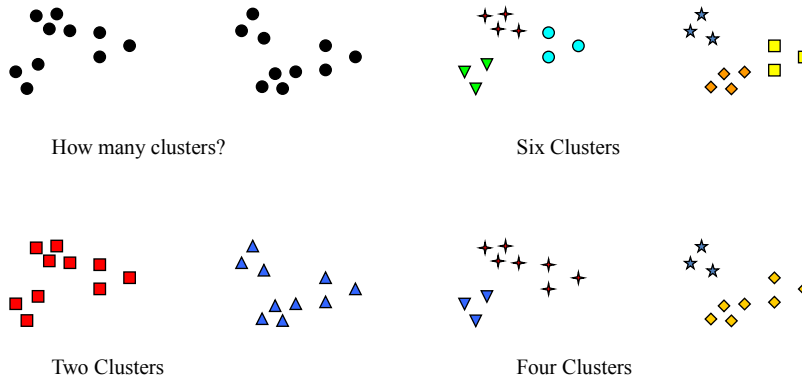
Other less glamorous things...

- Hospital Records
- Scientific Imaging
 - Related genes, related stars, related sequences
- Market Research
 - Segmenting markets, product positioning
- Social Network Analysis
- Data mining
- Image segmentation...

What is not Cluster Analysis?

- Supervised classification
 - Have class label information
- Simple segmentation
 - Dividing students into different registration groups alphabetically, by last name
- Results of a query
 - Groupings are a result of an external specification
- Graph partitioning
 - Some mutual relevance and synergy, but areas are not identical

Notion of a Cluster can be Ambiguous



Sajjad Haider

Fall 2012

9

Quality: What Is Good Clustering?

- A good clustering method will produce high quality clusters with
 - high intra-class similarity
 - low inter-class similarity
- The quality of a clustering result depends on both the similarity measure used by the method and its implementation
- The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns

Sajjad Haider

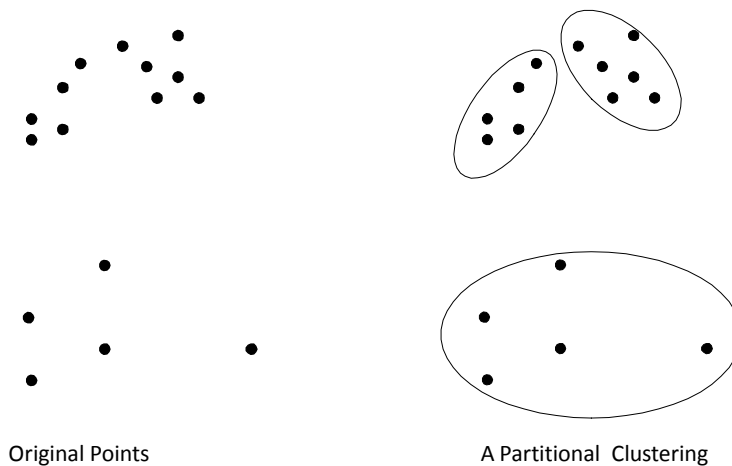
Fall 2012

10

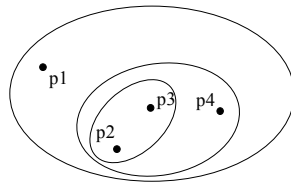
Types of Clusterings

- A **clustering** is a set of clusters
- Important distinction between **hierarchical** and **partitional** sets of clusters
- **Partitional Clustering**
 - A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset
- **Hierarchical clustering**
 - A set of nested clusters organized as a hierarchical tree

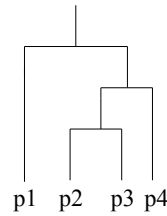
Partitional Clustering



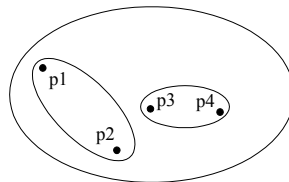
Hierarchical Clustering



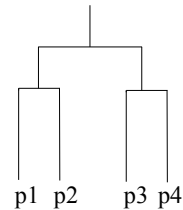
Traditional Hierarchical Clustering



Traditional Dendrogram



Non-traditional Hierarchical Clustering



Non-traditional Dendrogram

Sajjad Haider

Fall 2012

13

Measure the Quality of Clustering

- **Dissimilarity/Similarity metric:** Similarity is expressed in terms of a distance function, typically metric: $d(i, j)$
- There is a separate “quality” function that measures the “goodness” of a cluster.
- The definitions of **distance functions** are usually very different for interval-scaled, boolean, categorical, ordinal ratio, and vector variables.
- Weights should be associated with different variables based on applications and data semantics.
- It is hard to define “similar enough” or “good enough”
 - the answer is typically highly subjective.

Sajjad Haider

Fall 2012

14

What is Similarity?

The quality or state of being similar; likeness; resemblance; as, a similarity of features.

Webster's Dictionary



Similarity is hard to define, but...
"We know it when we see it"

The real meaning of similarity is a philosophical question. We will take a more pragmatic approach.

Sajjad Haider

Fall 2012

15

Partitioning Algorithms: Basic Concept

- Given a k , find a partition of k clusters that optimizes the chosen partitioning criterion
 - Global optimal: exhaustively enumerate all partitions
 - Heuristic methods: *k-means* and *k-medoids* algorithms
 - *k-means* (MacQueen'67): Each cluster is represented by the center of the cluster
 - *k-medoids* or PAM (Partition around medoids) (Kaufman & Rousseeuw'87): Each cluster is represented by one of the objects in the cluster

Sajjad Haider

Fall 2012

16

The *K-Means* Clustering Method

- Given k , the *k-means* algorithm is implemented in four steps:
 - Partition objects into k nonempty subsets
 - Compute seed points as the centroids of the clusters of the current partition (the centroid is the center, i.e., *mean point*, of the cluster)
 - Assign each object to the cluster with the nearest seed point
 - Go back to Step 2, stop when no more new assignment

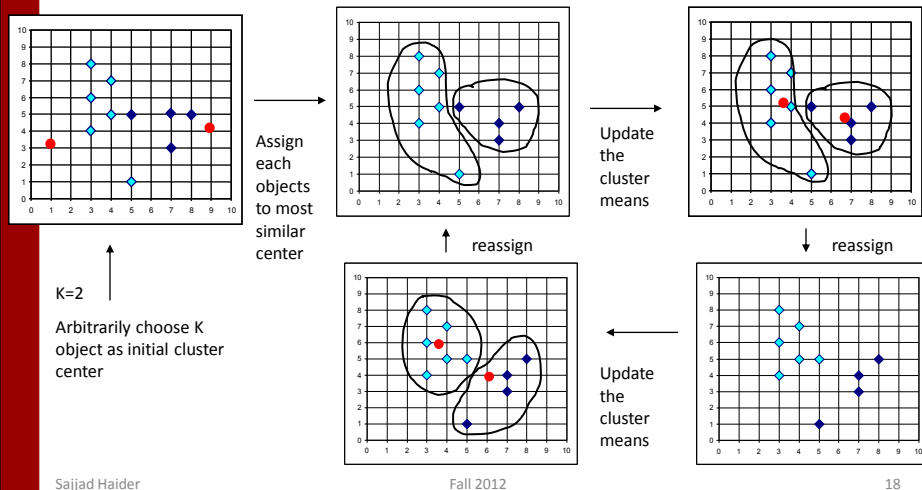
Sajjad Haider

Fall 2012

17

The *K-Means* Clustering Method (Cont'd)

- Example



Sajjad Haider

Fall 2012

18

The K-Means Algorithm

1. Choose a value for K, the total number of clusters to be determined.
2. Choose K instances within the dataset at random. These are the initial cluster centers.
3. Use simple Euclidean distance to assign the remaining instances to their closest cluster center.
4. Use the instance in each cluster to calculate a new mean for each cluster.
5. If the new mean values are identical to the mean values of the previous iteration the process terminates. Otherwise, use the new means as cluster centers and repeat steps 3-5.

Sajjad Haider

Fall 2012

19

Working of the K-Mean Algorithm

Instance #:	1	2	3	4	5	6
X:	1	1	2	2	3	3
Y:	1.5	4.5	1.5	3.5	2.5	6.0

- Let's pick Instances #1 and #3 as the initial centroids.

	Distance with Centroid1	Distance with Centroid2
• Instance #2	3.00	3.16
• Instance #4	2.24	2.00
• Instance #5	2.24	1.41
• Instance #6	6.02	5.41

- New centroids are (1, 3) and (2.5, 3.4)**

Sajjad Haider

Fall 2012

20

Comments on the *K-Means* Method

- Strength: *Relatively efficient*: $O(tkn)$, where n is # objects, k is # clusters, and t is # iterations. Normally, $k, t \ll n$.
- Comment: Often terminates at a *local optimum*. The *global optimum* may be found using techniques such as: *deterministic annealing* and *genetic algorithms*
- Weakness
 - Applicable only when *mean* is defined, then what about categorical data?
 - Need to specify k , the *number* of clusters, in advance
 - Unable to handle noisy data and *outliers*
 - Not suitable to discover clusters with *non-convex shapes*

Variations of the *K-Means* Method

- A few variants of the *k-means* which differ in
 - Selection of the initial k means
 - Dissimilarity calculations
 - Strategies to calculate cluster means
- Handling categorical data: *k-modes* (Huang'98)
 - Replacing means of clusters with modes
 - Using new dissimilarity measures to deal with categorical objects
 - Using a frequency-based method to update modes of clusters

K-Means Clustering – Details

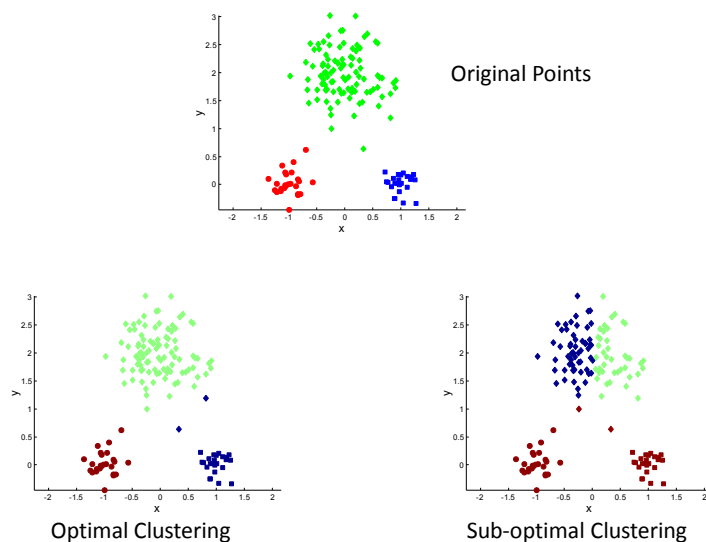
- Initial centroids are often chosen randomly.
 - Clusters produced vary from one run to another.
- The centroid is (typically) the mean of the points in the cluster.
- ‘Closeness’ is measured by Euclidean distance, cosine similarity, correlation, etc.
- K-means will converge for common similarity measures mentioned above.
- Most of the convergence happens in the first few iterations.
 - Often the stopping condition is changed to ‘Until relatively few points change clusters’

Sajjad Haider

Fall 2012

23

Two different K-means Clusterings

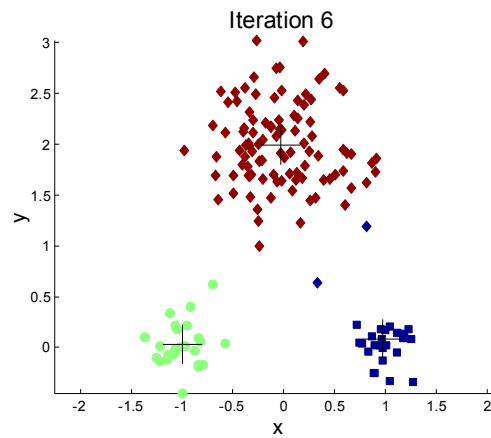


Sajjad Haider

Fall 2012

24

Importance of Choosing Initial Centroids

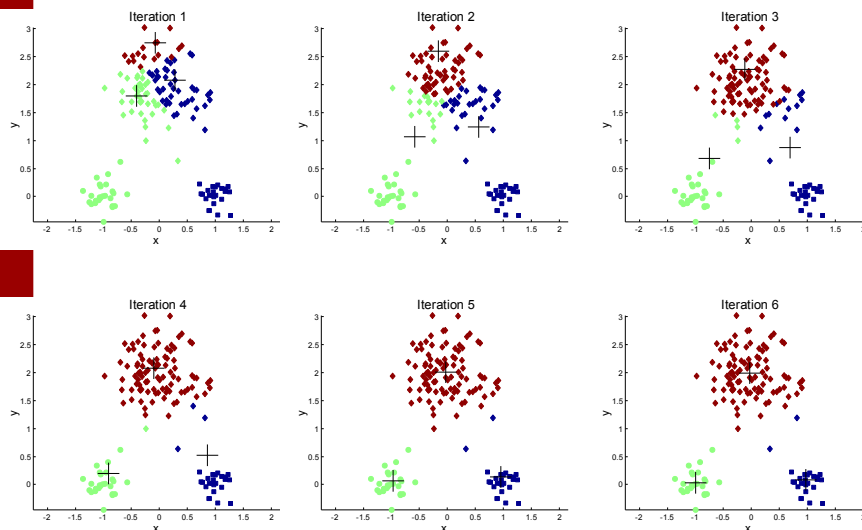


Sajjad Haider

Fall 2012

25

Importance of Choosing Initial Centroids

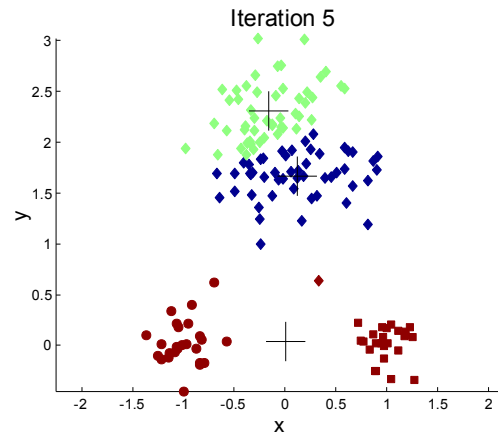


Sajjad Haider

Fall 2012

26

Importance of Choosing Initial Centroids ...

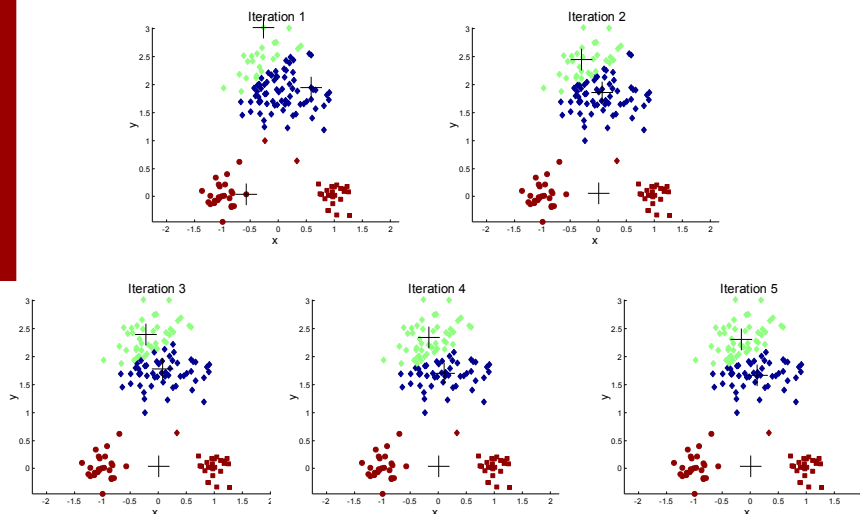


Sajjad Haider

Fall 2012

27

Importance of Choosing Initial Centroids ...



Sajjad Haider

Fall 2012

28

Evaluating K-means Clusters

- Most common measure is Sum of Squared Error (SSE)
 - For each point, the error is the distance to the nearest cluster
 - To get SSE, we square these errors and sum them.

$$SSE = \sum_{i=1}^K \sum_{x \in C_i} dist^2(m_i, x)$$

- x is a data point in cluster C_i and m_i is the representative point for cluster C_i
 - can show that m_i corresponds to the center (mean) of the cluster
- Given two clusters, we can choose the one with the smallest error
- One easy way to reduce SSE is to increase K , the number of clusters
 - A good clustering with smaller K can have a lower SSE than a poor clustering with higher K

Solutions to Initial Centroids Problem

- Multiple runs
 - Helps, but probability is not on your side
- Sample and use hierarchical clustering to determine initial centroids
- Select more than k initial centroids and then select among these initial centroids
 - Select most widely separated
- Postprocessing
- Bisecting K-means
 - Not as susceptible to initialization issues

Pre-processing and Post-processing

- Pre-processing
 - Normalize the data
 - Eliminate outliers
- Post-processing
 - Eliminate small clusters that may represent outliers
 - Split 'loose' clusters, i.e., clusters with relatively high SSE
 - Merge clusters that are 'close' and that have relatively low SSE

Bisecting K-means

- Bisecting K-means algorithm
 - Variant of K-means that can produce a partitional or a hierarchical clustering

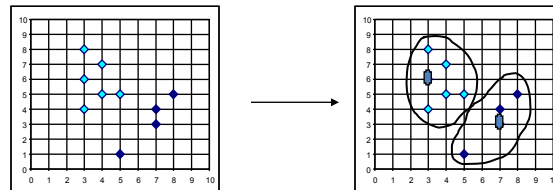
```

1: Initialize the list of clusters to contain the cluster containing all points.
2: repeat
3:   Select a cluster from the list of clusters
4:   for  $i = 1$  to number_of_iterations do
5:     Bisect the selected cluster using basic K-means
6:   end for
7:   Add the two clusters from the bisection with the lowest SSE to the list of clusters.
8: until Until the list of clusters contains  $K$  clusters

```

What Is the Problem of the K-Means Method?

- The k-means algorithm is sensitive to outliers !
 - Since an object with an extremely large value may substantially distort the distribution of the data.
- K-Medoids: Instead of taking the **mean** value of the object in a cluster as a reference point, **medoids** can be used, which is the **most centrally located** object in a cluster.



Sajjad Haider

Fall 2012

33

Limitations of K-means

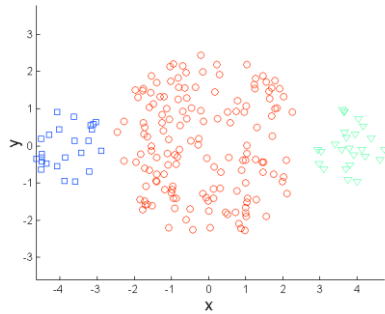
- K-means has problems when clusters are of differing
 - Sizes
 - Densities
 - Non-globular shapes
- K-means has problems when the data contains outliers.

Sajjad Haider

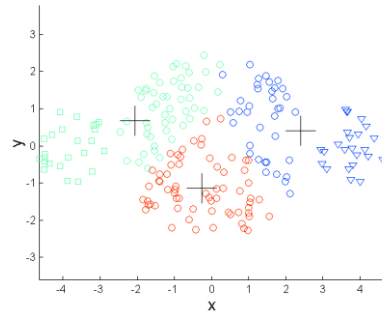
Fall 2012

34

Limitations of K-means: Differing Sizes



Original Points



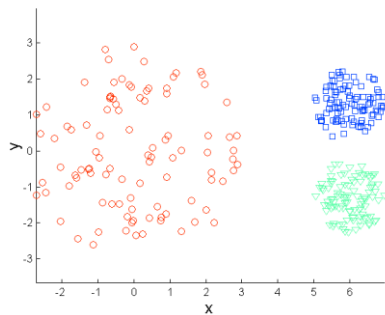
K-means (3 Clusters)

Sajjad Haider

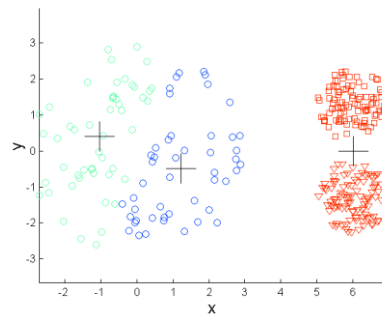
Fall 2012

35

Limitations of K-means: Differing Density



Original Points



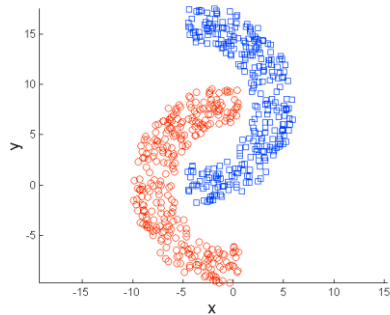
K-means (3 Clusters)

Sajjad Haider

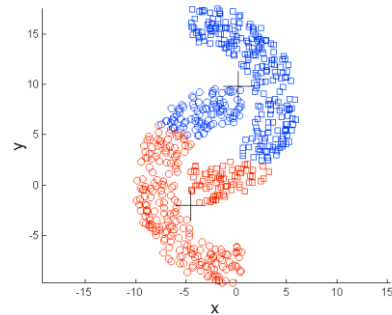
Fall 2012

36

Limitations of K-means: Non-globular Shapes



Original Points



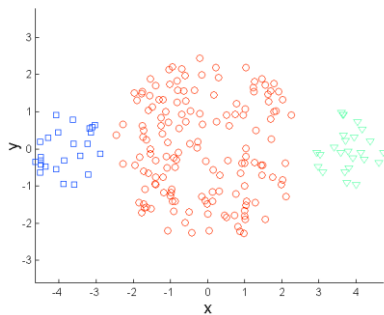
K-means (2 Clusters)

Sajjad Haider

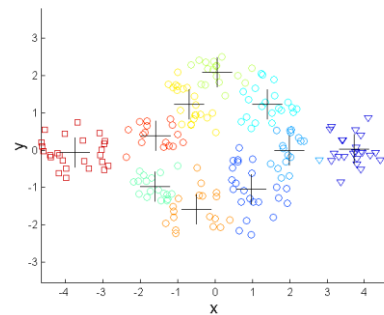
Fall 2012

37

Overcoming K-means Limitations



Original Points



K-means Clusters

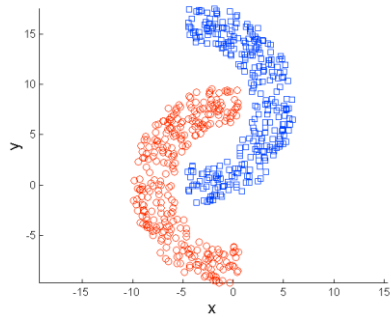
One solution is to use many clusters.
Find parts of clusters, but need to put together.

Sajjad Haider

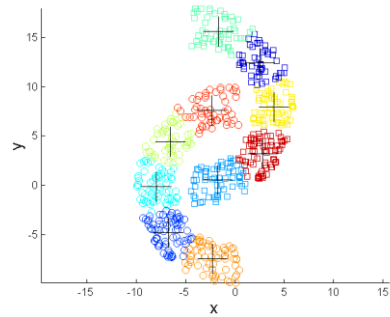
Fall 2012

38

Overcoming K-means Limitations



Original Points



K-means Clusters